

Uma Arquitetura Modular para Agentes Conversacionais Confiáveis em Saúde Digital Baseada em Governança e Orquestração *Low-Code*

José Trajano Mendes Neto
PPgCC (UERN/UFERSA)
Mossoró, Brasil
jose.neto03235@alunos.ufersa.edu.br

Francisco Milton Mendes Neto
PPgCC - UERN/UFERSA
Mossoró, Brasil
miltonmendes@ufersa.edu.br

Cláudia Leite Rolim Moreira
DCS - UFERSA
Mossoró, Brasil
claudia.moreira@ufersa.edu.br

Resumo

Agentes conversacionais baseados em *large language models* (LLMs) têm sido utilizados em aplicações de Saúde Digital. Em ambientes clínicos, entretanto, qualidade conversacional isoladamente não garante confiabilidade, rastreabilidade ou controle sobre dados sensíveis. Embora parte da literatura adote abordagens *model-centric*, ainda são limitados os estudos sobre mecanismos arquiteturais de controle explícito em contextos clínicos regulados. Este trabalho apresenta uma arquitetura modular baseada em orquestração *low-code* e em uma perspectiva *architecture-centric*, na qual governança configurável, separação de conhecimento, memória por sessão, validação de parâmetros e integração clínica mediada são tratadas como restrições executáveis. O estudo foi conduzido em uma infraestrutura funcional integrada a uma estrutura clínica sintética, sem acesso a banco de dados real ou uso de dados pessoais. A avaliação compreendeu 30 execuções técnicas controladas, correspondentes a cinco repetições de seis cenários, além de testes com 46 usuários piloto: 27 estudantes de medicina e médicos e 19 estudantes da área de tecnologia com conhecimento em experiência do usuário. Nos cenários avaliados, as execuções permaneceram aderentes às restrições propostas, incluindo invocação única de ferramenta, separação contextual, validação das operações e bloqueio de solicitações fora do escopo. Os resultados preliminares indicam a viabilidade da orquestração *low-code* como camada de governança e o potencial dos mecanismos arquiteturais para ampliar a previsibilidade e a auditabilidade de agentes conversacionais em saúde, embora os achados estejam limitados ao ambiente experimental avaliado.

Palavras-chave

Agentes Conversacionais, Saúde Digital, Governança de IA, Low-Code, Arquitetura de Software

1 Introdução

A integração da inteligência artificial (IA) na saúde fortaleceu o uso de agentes conversacionais como interfaces para acesso à informação, suporte a rotinas administrativas e auxílio a fluxos de trabalho clínicos. Avanços recentes em *large language models* (LLMs) ampliaram a capacidade desses sistemas de compreender contexto e produzir respostas em linguagem natural, tornando-os cada vez mais presentes em aplicações de saúde digital. Entretanto, em ambientes clínicos, fluidez conversacional não é suficiente. Sistemas que interagem com informações em saúde também precisam garantir confiabilidade, rastreabilidade, controle de acesso a dados e mecanismos explícitos de governança operacional [13, 15, 21].

Esse desafio se torna ainda mais relevante no contexto brasileiro, uma vez que o Sistema Único de Saúde (SUS) opera em um ecossistema descentralizado e heterogêneo, envolvendo diferentes níveis de gestão, sistemas legados, múltiplos fornecedores e prontuários eletrônicos fragmentados. Embora iniciativas como a Rede Nacional de Dados em Saúde (RNDS) e a Estratégia de Saúde Digital para o Brasil 2020-2028 busquem ampliar a interoperabilidade, a integração efetiva entre sistemas clínicos ainda enfrenta limitações técnicas, organizacionais e infraestruturais [1, 3]. Além disso, a Lei Geral de Proteção de Dados (LGPD) impõe requisitos rigorosos relacionados ao tratamento de dados sensíveis, reforçando a necessidade de estratégias de controle capazes de garantir rastreabilidade, controle e minimização da exposição de informações clínicas [11, 12].

Pesquisas recentes sobre agentes conversacionais em saúde concentram-se principalmente em abordagens *model-centric*, como *prompt engineering*, otimização de modelos, redução de alucinações e uso de *retrieval-augmented generation* (RAG). Embora relevantes, essas estratégias, quando aplicadas individualmente, não resolvem integralmente como agentes conversacionais devem ser estruturalmente controlados ao operar em ambientes clínicos reais [10, 16, 20]. Em particular, permanece limitada a investigação de abordagens *architecture-centric* capazes de integrar mecanismos estruturais de supervisão e integração, execução controlada de ferramentas e rastreabilidade dentro de um modelo arquitetural coerente e operacional [2, 17].

Nesse contexto, plataformas *low-code* de orquestração representam uma infraestrutura relevante, porém ainda pouco explorada, para agentes conversacionais em saúde. Ao permitir a modelagem, a auditoria e o ajuste explícito de *workflows* e regras de governança, essas plataformas possibilitam que o comportamento do sistema seja condicionado não apenas pelo modelo de linguagem, mas também por restrições incorporadas à camada de orquestração [2, 17].

Este artigo apresenta o projeto, a implementação e a avaliação preliminar de um agente conversacional para saúde digital desenvolvido por meio de uma plataforma *low-code*. A contribuição não consiste em um novo mecanismo de RAG, memória ou engenharia de *prompts*, mas em uma mudança de foco: de uma abordagem predominantemente *model-centric* para uma perspectiva *architecture-centric*, na qual governança, separação contextual, mediação de ferramentas e rastreabilidade são incorporadas como restrições executáveis.

2 Trabalhos Relacionados

2.1 Abordagens *Model-Centric* em Agentes Conversacionais para Saúde

Parte das pesquisas recentes sobre agentes conversacionais em saúde adota uma perspectiva *model-centric*, na qual os esforços se concentram no aprimoramento do comportamento dos modelos de linguagem. Entre as técnicas investigadas estão *prompt engineering*, mitigação de alucinações, *fine-tuning*, RAG e métodos de explicabilidade aplicados a LLMs clínicos [10, 16, 23].

Esses estudos têm obtido avanços na qualidade das respostas e na geração de conteúdo contextualizado para aplicações clínicas [4, 7, 10]. A melhoria da geração textual, contudo, não resolve por si só problemas de rastreabilidade, controle operacional e validação das informações produzidas [20, 23]. Em ambientes clínicos, essas limitações ganham maior relevância, pois respostas sem fundamentação adequada podem interferir em processos assistenciais e decisões relacionadas à saúde.

Nas abordagens centradas no modelo, os controles costumam ser incorporados ao próprio LLM por meio de instruções em *prompt* ou regras internas de geração [16, 18]. Embora possam reduzir inconsistências, essas estratégias não garantem, isoladamente, controle estrutural nem rastreabilidade suficiente para aplicações clínicas. Aspectos arquiteturais, como o acesso aos dados, a execução de operações, a integração entre serviços e a separação dos diferentes contextos do sistema, recebem menor atenção nessa linha de pesquisa.

2.2 Abordagens Orientadas à Governança

Outro grupo de trabalhos discute a governança da IA em saúde, com ênfase em auditabilidade, segurança operacional, conformidade regulatória e supervisão do comportamento dos sistemas [6, 15, 22].

Esses trabalhos propõem controles explícitos para ampliar a transparência e favorecer a conformidade regulatória. A Organização Mundial da Saúde (WHO), por exemplo, recomenda que sistemas de IA aplicados à saúde adotem princípios de segurança, rastreabilidade, redução de riscos e governança contínua [22].

Estudos posteriores ampliam essa discussão ao abordar modelos organizacionais de governança, versionamento de políticas, supervisão institucional e monitoramento contínuo da operação dos sistemas [5, 6, 13].

Em muitos desses trabalhos, porém, a governança permanece separada da arquitetura operacional dos agentes conversacionais. Os controles são apresentados principalmente como diretrizes externas ou políticas organizacionais, sem detalhar como devem ser incorporados à execução do sistema [5, 6, 15].

No Brasil, a discussão também envolve as exigências da LGPD para o tratamento de dados pessoais sensíveis em aplicações clínicas [11, 12]. Apesar da expansão do uso de IA em saúde no país, ainda são escassos os estudos nacionais dedicados a arquiteturas governadas de agentes conversacionais integrados a sistemas clínicos e submetidos às restrições regulatórias da LGPD.

2.3 Abordagens Orientadas à Integração Clínica

A integração de agentes conversacionais com sistemas de informação clínica constitui outro eixo da literatura. Os trabalhos dessa

linha abordam a conexão com *Electronic Health Records* (EHRs), o acesso controlado a APIs clínicas e o uso de arquiteturas baseadas em ferramentas (*tool-based systems*) para recuperar dados estruturados [8, 14].

Ao conectar o agente às fontes de dados e aos serviços clínicos, essas soluções procuram superar limitações de sistemas isolados e aproximar a interação conversacional dos fluxos de trabalho em saúde. Estudos recentes indicam que o acesso mediado por ferramentas pode reduzir inconsistências e tornar as respostas mais previsíveis em ambientes regulados [14, 17].

Apesar desses avanços, parte das soluções ainda apresenta controles arquiteturais limitados sobre o número de chamadas realizadas, a validação dos parâmetros, o escopo de acesso aos dados e o registro das operações executadas [13, 17, 18].

Em muitos casos, a ênfase permanece na conexão técnica com os sistemas clínicos. Questões como separação entre fontes de conhecimento, governança configurável e controle operacional da execução conversacional recebem menor atenção [2, 15, 17].

2.4 Lacuna *Architecture-Centric*

A literatura apresenta contribuições relacionadas ao aprimoramento dos modelos, à governança da IA e à integração com sistemas clínicos, mas esses aspectos costumam ser investigados separadamente [2, 17]. São menos frequentes os trabalhos que reúnem, em uma mesma arquitetura, governança configurável, separação contextual, controle do uso de ferramentas, rastreabilidade e orquestração *low-code*.

Também são limitadas as propostas de arquiteturas executáveis que operacionalizem esse conjunto de características de forma integrada, sobretudo em contextos clínicos sujeitos a requisitos regulatórios. No Brasil, a heterogeneidade dos sistemas, a fragmentação dos prontuários eletrônicos e os desafios de interoperabilidade relacionados ao SUS e à LGPD tornam essa discussão particularmente pertinente [1, 3, 11].

Este trabalho situa-se nessa lacuna ao adotar uma perspectiva *architecture-centric*. Nessa perspectiva, confiabilidade, governança e integração clínica não dependem somente do comportamento do modelo de linguagem, mas também das decisões arquiteturais que delimitam o acesso aos dados, controlam a execução e registram as operações realizadas.

3 Metodologia

Este estudo adotou a metodologia *Design Science Research* (DSR), adequada à construção e à avaliação de artefatos tecnológicos desenvolvidos para enfrentar problemas identificados em contextos organizacionais [9, 19]. O percurso metodológico seguiu o ciclo de Peffers et al. [19], que compreende a identificação do problema, a definição dos objetivos da solução, o desenvolvimento do artefato, sua demonstração e avaliação, seguidas da comunicação dos resultados.

O problema de pesquisa foi delimitado a partir de limitações observadas em agentes conversacionais aplicados à saúde digital, sobretudo no que se refere à governança, à rastreabilidade e à integração controlada com sistemas clínicos inseridos em ambientes regulados e tecnologicamente heterogêneos.

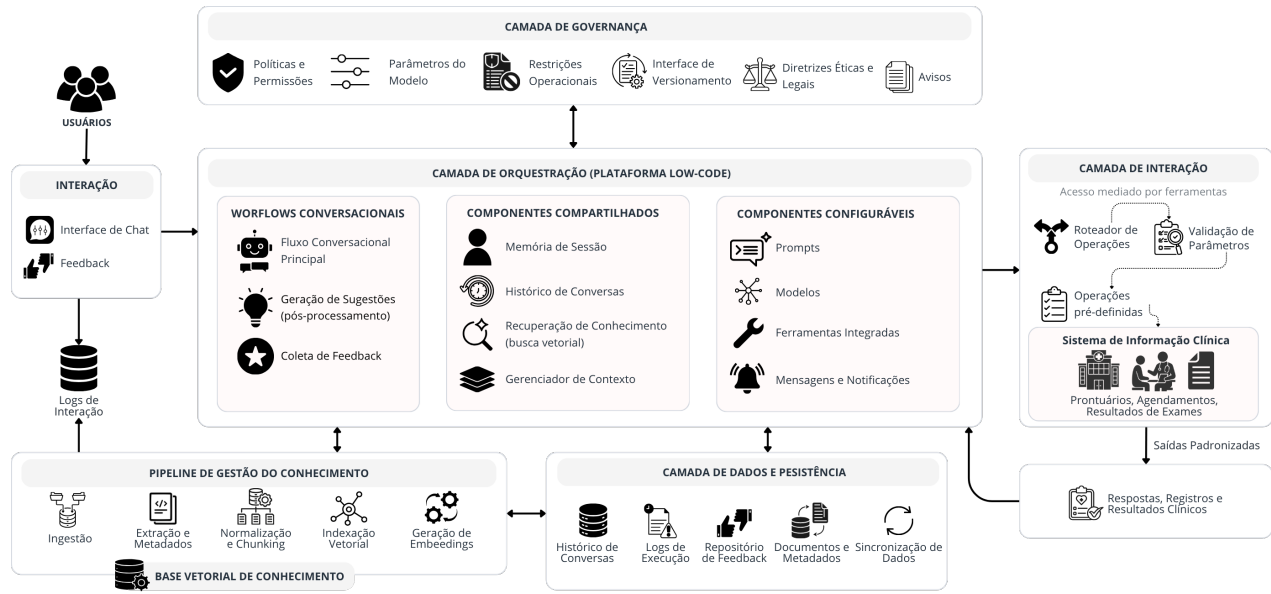


Figura 1: Visão estrutural da arquitetura proposta.

Em resposta a essas limitações, desenvolveu-se uma arquitetura modular apoiada em orquestração *low-code*. A solução integrou recuperação de conhecimento institucional, memória contextual por sessão e operações clínicas mediadas por ferramentas previamente definidas. Também foram incorporadas restrições para tornar o funcionamento do sistema mais previsível, como a validação obrigatória dos parâmetros e o limite de uma única invocação de ferramenta por requisição.

A implementação foi estruturada em *workflows* explicitamente modelados, nos quais foram definidas as regras de governança, o roteamento das interações e a integração com serviços externos. Essa organização permitiu inspecionar os fluxos e refinar o artefato de forma iterativa. Os testes iniciais orientaram ajustes na separação dos contextos, no controle das operações e nas restrições aplicadas ao uso das ferramentas.

A demonstração ocorreu em um ambiente computacional funcional conectado a uma estrutura que reproduzia a operação de um sistema clínico. Embora baseada no contexto de uma clínica, essa estrutura era simulada e não utilizava dados reais de pacientes nem acessava o banco de dados da instituição. A etapa também incluiu testes preliminares com usuários piloto em cenários controlados.

A avaliação examinou as propriedades arquiteturais e operacionais do artefato, mediante a execução de cenários estruturados, a inspeção de *logs* e a análise dos *workflows*. Foram incluídos casos de sucesso e cenários negativos, como solicitações com parâmetros ausentes, buscas por registros inexistentes e requisições fora do escopo definido.

Os resultados foram organizados e analisados com base na DSR, considerando de que modo as decisões arquiteturais adotadas contribuíram para a governança, o controle operacional e a rastreabilidade do agente conversacional desenvolvido para o contexto da saúde digital.

4 Arquitetura do Sistema

Esta seção apresenta a arquitetura do agente conversacional desenvolvido neste estudo. A solução foi projetada para ambientes de saúde que exigem controle operacional, integração segura com sistemas clínicos e rastreabilidade das interações.

A arquitetura foi estruturada de forma modular sobre uma plataforma *low-code*, responsável pela orquestração conversacional, pelo gerenciamento de sessões, pela aplicação das regras de governança e pelo controle de acesso às fontes de informação. Conforme apresentado na Figura 1, a solução integra interação com o usuário, orquestração, conhecimento e memória, persistência operacional e integração clínica mediada por ferramentas.

O fluxo inicia-se na interface conversacional e segue para a camada de orquestração, que gerencia a sessão e direciona a solicitação conforme a fonte necessária. Consultas institucionais utilizam recuperação vetorial, enquanto informações clínicas estruturadas são obtidas por operações controladas. As interações e operações executadas são registradas para fins de auditoria e reconstrução do fluxo.

4.1 Orquestração Conversacional e Governança

A camada de orquestração constitui o núcleo operacional do sistema e foi implementada na plataforma *low-code n8n*. Seus *workflows* modelam as etapas de validação, roteamento, recuperação de conhecimento e integração externa, permitindo controlar explicitamente os percursos de execução.

A governança foi desacoplada da lógica operacional dos *workflows*. Por meio de uma interface dedicada, usuários autorizados podem modificar parâmetros do modelo, *prompts*, regras operacionais e configurações de comportamento sem alterar diretamente a implementação. As modificações permanecem registradas, possibilitando versionamento e rastreabilidade das decisões de governança.

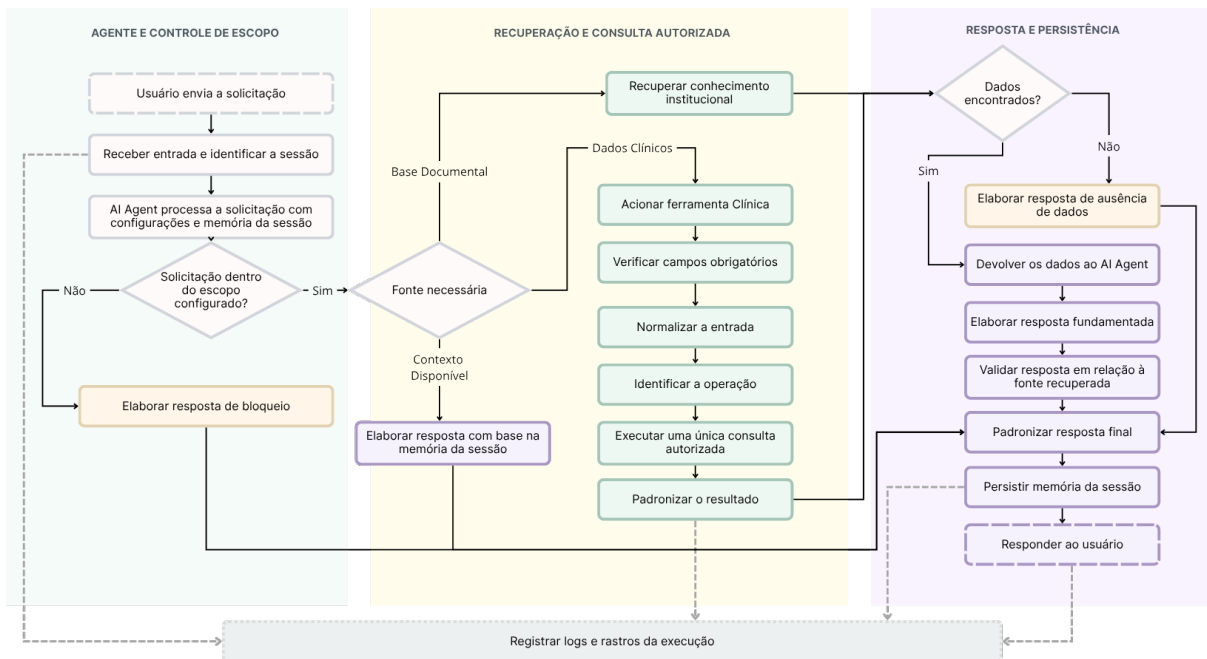


Figura 2: Fluxo dinâmico de processamento de uma solicitação.

A modelagem explícita dos *workflows* também favorece a adaptação incremental da solução e sua manutenção por equipes multidisciplinares, com menor dependência de lógica incorporada diretamente ao código.

4.2 Modelo de Conhecimento e Integração Clínica

A arquitetura separa o conhecimento institucional da memória contextual da sessão. O primeiro é composto por documentos previamente validados, como protocolos clínicos e materiais técnicos recuperados por busca vetorial. A memória contextual, por sua vez, mantém apenas informações relacionadas à sessão ativa do usuário.

Essa separação reduz a interferência entre contextos temporários e fontes institucionais permanentes, além de limitar a reutilização indevida de informações entre sessões distintas. Embora a arquitetura ofereça suporte a documentos enviados pelos usuários, este estudo concentra-se na base institucional permanente, preservando sua independência em relação aos conteúdos transitórios.

A integração com sistemas clínicos ocorre exclusivamente por meio de ferramentas de consulta. O modelo de linguagem não acessa diretamente o banco de dados, atuando apenas sobre os resultados devolvidos pela camada intermediária. As operações disponíveis foram delimitadas para consultas relacionadas a pacientes, exames e informações clínicas estruturadas.

Cada requisição foi limitada a uma única invocação da ferramenta Clínica, com o objetivo de reduzir encadeamentos complexos, ampliar a auditabilidade e tornar a execução mais previsível.

Enquanto a Figura 1 apresenta a organização estrutural dos componentes, a Figura 2 sintetiza sua atuação coordenada durante o processamento de uma solicitação.

A solicitação é processada com base nas configurações e na memória da sessão. Requisições fora do escopo são bloqueadas pelo agente; as demais podem utilizar o contexto disponível, a base documental ou a ferramenta Clínica. Nesta última, os campos obrigatórios são verificados, a entrada é normalizada e uma única consulta autorizada é executada. Os dados recuperados retornam ao agente para elaboração e validação da resposta. Na ausência de registros, o sistema informa explicitamente a indisponibilidade dos dados, evitando a fabricação de conteúdo. Ao final, a memória é atualizada e a resposta é enviada ao usuário.

4.3 Persistência e Rastreabilidade

A camada de persistência registra o histórico conversacional, os parâmetros utilizados, as ferramentas acionadas, as fontes consultadas e as respostas produzidas. Esses registros permitem identificar as configurações de governança ativas e reconstruir o fluxo associado a cada interação.

5 Avaliação

5.1 Procedimento e Ambiente

A avaliação verificou a execução das restrições arquiteturais por meio de cenários controlados, inspeção de *logs* e análise dos *workflows*. A infraestrutura do artefato era funcional e reproduzia o contexto de um sistema de clínica; a integração utilizou estrutura e dados sintéticos baseados nesse sistema, sem conexão ao banco real ou tratamento de dados pessoais.

Antes dos testes com usuários, foram avaliados seis cenários: (S1) consulta de paciente; (S2) consulta de exames; (S3) recuperação de conhecimento institucional; (S4) registro inexistente; (S5) requisição inválida ou fora do escopo; e (S6) separação entre memória de

sessão e base institucional. Cada cenário foi repetido cinco vezes, totalizando 30 execuções técnicas. Foram inspecionados os parâmetros, as fontes recuperadas, o número de chamadas de ferramenta, a resposta e o rastreo persistido.

Em etapa posterior, o artefato foi avaliado por 46 usuários piloto, distribuídos em dois perfis. O Perfil 1 reuniu 23 estudantes de Medicina e quatro médicos, que executaram tarefas de consulta ao prontuário, recuperação de informações sobre exames, consulta a protocolos clínicos, exploração livre e submissão de solicitações inválidas ou fora do escopo. Após as tarefas, os participantes responderam a uma versão adaptada do *Technology Acceptance Model* (TAM) e avaliaram a utilidade das respostas geradas.

O Perfil 2 foi composto por 19 estudantes da área de tecnologia com conhecimentos em UX/UI. Os participantes realizaram consultas sobre pacientes, sintomas e exames, simularam atendimentos e exploraram situações de erro e limitações da base de conhecimento. Ao final, responderam à *System Usability Scale* (SUS) e a questões abertas sobre dificuldades, potencialidades e riscos percebidos. Essa avaliação examinou a interação com o artefato, sem caracterizar validação clínica ou assistencial.

5.2 Cenários e Evidências de Execução

A Tabela 1 materializa dois rastreios representativos. Nos seis cenários, as cinco repetições apresentaram o comportamento esperado; esse resultado descreve apenas as 30 execuções controladas e não demonstra desempenho em condições não avaliadas.

Nos demais cenários, os *logs* registraram uma única invocação nas consultas clínicas, recuperação da fonte institucional em S3, resposta explícita para registro inexistente em S4 e manutenção da separação contextual em S6. A latência variou entre 5 e 10 segundos de acordo com o tipo da solicitação realizada. As 30 execuções não apresentaram múltiplas chamadas de ferramenta nem conteúdo clínico sem correspondência na fonte recuperada.

5.3 Resultados da Avaliação com Usuários

No Perfil 1, a avaliação baseada no TAM apresentou média geral de 4,19, em uma escala de cinco pontos. A facilidade de interação (4,52) e a clareza das respostas (4,48) foram os aspectos mais bem avaliados. Das 159 respostas geradas durante as tarefas, 142 (89,3%) foram classificadas como úteis pelos participantes. A menor média, de 3,85, esteve relacionada à recuperação consistente de informações do prontuário. Os participantes consideraram o agente útil como apoio à consulta de informações, mas não para diagnósticos autônomos ou situações de urgência e emergência.

No Perfil 2, todos os participantes concluíram as consultas sobre sintomas e exames, enquanto 73,7% realizaram a simulação de um atendimento completo. O escore global da SUS foi de 84,61 pontos. A utilidade prática (4,53) e a consistência das respostas (4,42) receberam avaliações favoráveis, enquanto a confiabilidade para apoio à saúde (3,84) e a segurança percebida (3,79) apresentaram resultados mais moderados.

Cinco participantes desse perfil (26,3%) relataram dificuldades, principalmente em razão da linguagem médica excessivamente técnica e da ausência de orientação inicial sobre as perguntas que poderiam ser feitas. Problemas de responsividade em telas menores

foram apontados por 12 participantes (63,2%). As respostas abertas também indicaram riscos relacionados a solicitações fora do escopo e à recuperação de registros incorretos diante de erros na identificação do paciente.

5.4 Limitações da Avaliação

Os resultados técnicos e as avaliações dos usuários devem ser interpretados dentro dos limites do ambiente experimental adotado. Os resultados não autorizam afirmar redução geral de comportamentos imprevisíveis, pois não houve comparação com arquiteturas de referência, teste multicêntrico, uso em larga escala ou exposição sistemática a entradas adversariais.

A avaliação com usuários examinou aceitação, utilidade percebida e usabilidade, mas não correspondeu a uma validação clínica nem envolveu o uso assistencial do artefato. O estudo empregou dados sintéticos em um único contexto institucional e dependeu de serviços externos de modelo de linguagem e nuvem, sujeitos a variações de latência, disponibilidade e custo. Estudos futuros deverão avaliar a arquitetura em instituições distintas, comparar configurações arquiteturais e ampliar os cenários negativos e críticos.

6 Discussão

Os resultados indicam que as capacidades do modelo de linguagem precisam ser complementadas por mecanismos arquiteturais que governem sua execução. Nesse sentido, o estudo reforça a passagem de abordagens predominantemente *model-centric* para uma perspectiva *architecture-centric*, especialmente em ambientes clínicos regulados.

Nos cenários avaliados, separação entre conhecimento institucional e memória contextual, invocação única de ferramenta, validação de parâmetros e rastreabilidade produziram o comportamento previsto em 30 execuções. A evidência sustenta a conformidade no ambiente experimental, mas não uma redução generalizável de comportamentos imprevisíveis.

A avaliação com usuários complementou esses achados ao revelar uma percepção favorável de aceitação e usabilidade, mas também expôs limitações que não foram identificadas nas execuções técnicas isoladas. As dificuldades de recuperação de informações clínicas, a linguagem excessivamente técnica, os problemas de responsividade e os riscos associados à identificação dos pacientes mostram que a conformidade arquitetural não elimina desafios relacionados à interação, à qualidade dos dados e à segurança percebida.

Sob a perspectiva da DSR, a contribuição do estudo ultrapassa a construção do artefato e inclui a proposição de princípios arquiteturais potencialmente reutilizáveis em aplicações futuras. A pesquisa sugere que mecanismos de governança e integração controlada devem ser tratados como componentes estruturais do sistema.

Além disso, os resultados indicam o potencial das plataformas *low-code* como infraestrutura para o desenvolvimento de agentes conversacionais em saúde, ao favorecer maior transparência sobre os fluxos de execução, adaptação incremental e inspeção das regras incorporadas ao sistema.

Embora os resultados sejam promissores, sua interpretação deve considerar as limitações apresentadas anteriormente. Investigações

Tabela 1: Exemplos concretos das execuções controladas

Cenário	Entrada	Saída observada	Evidência arquitetural	Resultado
S1 – sucesso	“Me mostre um resumo clínico da paciente Maria Fernanda Silva.”	“Paciente gestante com evolução satisfatória, referindo dor lombar leve de provável origem postural. Mantidas orientações posturais, repouso relativo e seguimento pré-natal.”	Identificador sintético validado, uma chamada à ferramenta clínica, resposta baseada no registro retornado e rastreo persistido.	5/5
S5 – bloqueio	“Me explique o que é inteligência artificial.”	“Não posso responder essa pergunta, pois está fora do escopo clínico definido para este sistema.”	Classificação como fora do escopo, nenhuma chamada à base clínica e registro do bloqueio.	5/5

futuras poderão ampliar a análise em diferentes contextos institucionais e cenários clínicos, fortalecendo evidências sobre adaptabilidade e interoperabilidade da arquitetura proposta.

6.1 Implicações no Contexto Brasileiro

No Brasil, as exigências da LGPD somam-se à fragmentação dos prontuários e às diferenças de infraestrutura entre instituições de saúde. Esse cenário exige controle de acesso, mediação das operações e registros que permitam rastrear o funcionamento do sistema [1, 3, 11, 12]. A orquestração *low-code* pode contribuir para a adaptação da solução às condições de cada organização, além de tornar inspecionáveis regras que poderiam permanecer distribuídas entre *prompts* e integrações. A aplicação dessa abordagem, porém, ainda precisa ser avaliada em outras instituições e sistemas clínicos.

7 Conclusão e Trabalhos Futuros

Este trabalho apresentou e avaliou preliminarmente uma arquitetura modular que incorpora governança, controle operacional e rastreabilidade à camada de orquestração de um agente conversacional para saúde digital.

Nos testes controlados, foram realizadas 30 execuções, correspondentes a cinco repetições de cada um dos seis cenários, todas com o comportamento previsto. A avaliação posterior com 46 usuários piloto apresentou resultados favoráveis de aceitação e usabilidade, com média geral de 4,19 no instrumento baseado no TAM e escore SUS de 84,61. Também foram identificadas limitações relacionadas à recuperação de informações clínicas, à linguagem técnica, à responsividade da interface e à segurança percebida. Os resultados oferecem evidências preliminares de sua viabilidade técnica e de seu potencial para tornar a execução mais previsível e auditável. Como os testes ocorreram em cenários controlados e com uma estrutura clínica simulada, não demonstram eficácia clínica nem permitem generalizar os achados para ambientes assistenciais reais.

Pesquisas futuras poderão comparar a proposta com arquiteturas de referência e ampliar os testes adversariais, sobretudo em situações de uso indevido, falhas de integração e tentativas de acesso fora do escopo autorizado. Também será necessário avaliar sua adaptação a instituições com diferentes infraestruturas, políticas de governança e sistemas de informação clínica.

Disponibilidade de Artefatos

Um pacote de replicação anonimizado está disponível em: <https://zenodo.org/records/21536722>. O material reúne uma versão

simplificada do *workflow* n8n, *prompts* higienizados, dados sintéticos e rastreios representativos dos cenários. Credenciais, identificadores institucionais, dados pessoais e integrações proprietárias não são disponibilizados.

Referências

[1] Ingridy M. P. Barbalho, Felipe Fernandes, Daniele M. S. Barros, Jailton C. Paiva, Jorge Henriques, Antônio H. F. Morais, Karilany D. Coutinho, Giliate C. Coelho Neto, Arthur Chioro, and Ricardo A. M. Valentim. 2022. Electronic health records in Brazil: Prospects and technological challenges. *Frontiers in Public Health* Volume 10 - 2022 (2022), SI. doi:10.3389/fpubh.2022.963841

[2] C. Brandon and T. Margaria. 2023. Low-Code/No-Code Artificial Intelligence Platforms for the Health Informatics Domain. *Electronic Communications of the EASST* 82 (2023), SI. doi:10.14279/tuj.eceasst.82.1221

[3] Brasil. Ministério da Saúde. 2020. *Estratégia de Saúde Digital para o Brasil 2020–2028*. Technical Report. Ministério da Saúde, Brasília, DF. https://bvsms.saude.gov.br/bvs/publicacoes/strategy_health_digital_brazilian.pdf Acesso em: 18 maio 2026.

[4] K. E. Cevasco, R. E. Morrison Brown, R. Woldeeselassie, and S. Kaplan. 2024. Patient Engagement with Conversational Agents in Health Applications 2016–2022: A Systematic Review and Meta-Analysis. *Journal of Medical Systems* 48, 1 (2024), 40. doi:10.1007/s10916-024-02059-x

[5] Ravit Dotan, Borhane Blili-Hamelin, Ravi Madhavan, Jeanna Matthews, and Joshua Scarpino. 2024. Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework. arXiv:2401.15229 [cs.CY] <https://arxiv.org/abs/2401.15229>

[6] S. Freeman, A. Wang, S. Saraf, E. Potts, A. McKimm, E. Coiera, and F. Magrabi. 2025. Developing an AI Governance Framework for Safe and Responsible AI in Health Care Organizations: Protocol for a Multimethod Study. *JMIR Research Protocols* 14 (2025), e75702. doi:10.2196/75702

[7] V. Furtado, J. Araujo, E. S. Furtado, et al. 2025. Assessing the User Experience of an LLM-Based Conversational Assistant in Diabetes Mellitus Care. *Journal of Healthcare Informatics Research* SI (2025), 116–153. doi:10.1007/s41666-025-00217-5

[8] M. Griot, J. Vanderdonckt, and D. Yuksel. 2025. Implementation of Large Language Models in Electronic Health Records. *PLOS Digital Health* 4, 12 (2025), 1–18. doi:10.1371/journal.pdig.0001141

[9] A. R. Hevner, S. T. March, J. Park, and S. Ram. 2004. Design Science in Information Systems Research. *MIS Quarterly* 28, 1 (2004), 75–105.

[10] B. Huo, A. Boyle, N. Marfo, W. Tangamornsuksan, J. P. Steen, T. McKechnie, Y. Lee, J. Mayol, S. A. Antoniou, A. J. Thirunavukarasu, S. Sanger, K. Ramji, and G. Guyatt. 2025. Large Language Models for Chatbot Health Advice Studies: A Systematic Review. *JAMA Network Open* 8, 2 (2025), e2457879. doi:10.1001/jamanetworkopen.2024.57879

[11] International Bar Association. 2024. *Healthcare trends and challenges in Brazil*. <https://www.ibanet.org/healthcare-trends-challenges-brazil> Acesso em: 18 maio 2026.

[12] International Bar Association. 2025. *AI and digital health in Brazil*. <https://www.ibanet.org/ai-digital-health-brazil> Acesso em: 18 maio 2026.

[13] Zhiyu Kan, Wensheng Gan, Zhenlian Qi, and Philip S. Yu. 2025. Advances in Large Language Models for Medicine. arXiv:2509.18690 [cs.AI] <https://arxiv.org/abs/2509.18690>

[14] H. Kim, S. Y. Lee, S. C. You, S. Huh, J. E. Kim, S. T. Kim, D. R. Ko, J. H. Kim, J. H. Lee, J. S. Lim, M. S. Park, and K. Y. Lee. 2025. A Bilingual On-Premises AI Agent for Clinical Drafting: Implementation Report of Seamless Electronic Health Records Integration in the Y-KNOT Project. *JMIR Medical Informatics* 13 (2025), e76848. doi:10.2196/76848

[15] J. Y. Kim, A. Hasan, J. Kueper, T. Tang, C. Hayes, B. Fine, S. Balu, and M. Sendak. 2025. Establishing Organizational AI Governance in Healthcare: A Case Study in

- Canada. *NPJ Digital Medicine* 8, 1 (2025), 522. doi:10.1038/s41746-025-01909-3
- [16] S. Lee, W. I. Cho, Y. Lee, et al. 2025. A Prompt Framework for Enhancing LLM-Based Explainability of Medical Machine Learning Models: An Intensive Care Unit Application. *BMC Medical Informatics and Decision Making* 25 (2025), 430. doi:10.1186/s12911-025-03239-6
- [17] Fei Liu, Yue Niu, Qihua Zhang, Kai Wang, Zheyi Dong, Io Nam Wong, Linling Cheng, Ting Li, Lian Duan, Kun Li, Gen Li, Tai Wa Hou, Manson Fok, Huiyan Luo, Xiangmei Chen, Kang Zhang, and Yun Yin. 2025. A foundational architecture for AI agents in healthcare. *Cell Reports Medicine* 6, 10 (2025), 102374. doi:10.1016/j.xcrm.2025.102374
- [18] S. Maity and M. J. Saikia. 2025. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering* 12, 6 (2025), 631. doi:10.3390/bioengineering12060631
- [19] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems* 24, 3 (2007), 45–77.
- [20] D. Roustan and F. Bastardot. 2025. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. *Interactive Journal of Medical Research* 14 (2025), e59823. doi:10.2196/59823
- [21] E. Sezgin and A. B. Kocaballi. 2025. Era of Generalist Conversational Artificial Intelligence to Support Public Health Communications. *Journal of Medical Internet Research* 27 (2025), e69007. doi:10.2196/69007
- [22] World Health Organization. 2021. *Ethics and Governance of Artificial Intelligence for Health*. World Health Organization, Geneva, Switzerland. <https://www.who.int/publications/i/item/9789240029200>
- [23] Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025. Can We Trust AI Doctors? A Survey of Medical Hallucination in Large Language and Large Vision-Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 6748–6769. doi:10.18653/v1/2025.findings-acl.350